Addressing the Technical Friction Points in Open Science

Jonathan A. Rees
Creative Commons
15 November 2011

Introduction

Science, in principle, is characterized by exposure to scrutiny of all aspects of an investigation - assumptions, inputs, observations, experiments, and analysis. It is from transparency that all the benefits of scientific method spring: verifiability, reproducibility, accountability, and the chance to improve or build on what has been done. In a sense the phrase “open science” is redundant because a study that is not adequately open is at risk for being is noun scientific.

But there are degrees of transparency. A study whose exposition and explanations are difficult to obtain or expensive to process imposes transaction costs - mainly in terms of researcher time - on those who would want to examine it. Chasing down literature and other inputs, reproduction of experiment or analysis, new analysis of its data, meta-analysis and comparisons, all take time and attention.

If appropriate techniques and tools can be brought to bear at each end of the transaction such tasks can be made more efficient. Prior to the advent of the internet, communication and distribution costs, in practice, dominated transaction costs. The internet virtually eliminates these costs and opens up the possibility of techniques and tools that were previously impossible. At the scale at which science is practiced now, the benefits of consequent efficiencies could be staggering.

By placing so many scientific resources at hand, the Internet brings transaction costs into relief as each investigator attempts to make the best possible use of all available resources. Scaling issues are coming to dominate research practice: how many resources can an investigator process, and how quickly, whether for search, aggregation, reanalysis, or any other purpose?

Technical approaches to reducing transaction costs require that the resources generated at one step in the chain (a) are amenable to automated processing, (b) can be used with the tools that will be available to those receiving them.

The cost of using a resource is highly variable. We have the following rough hierarchy

- resource not shared, or prohibitive $ cost
  - resource available, but incomprehensible
    - interpreting the resource requires intelligence or judgment - human in the loop
      - interpreting the resource requires complex, expensive software, or heuristics
        - interpreting the resource requires relatively simple, rigorous software

with transaction cost decreasing (interoperability increasing) as we go down the list. (Of course several levels may apply simultaneously, because the nature of the use affects the kind of processing required.)

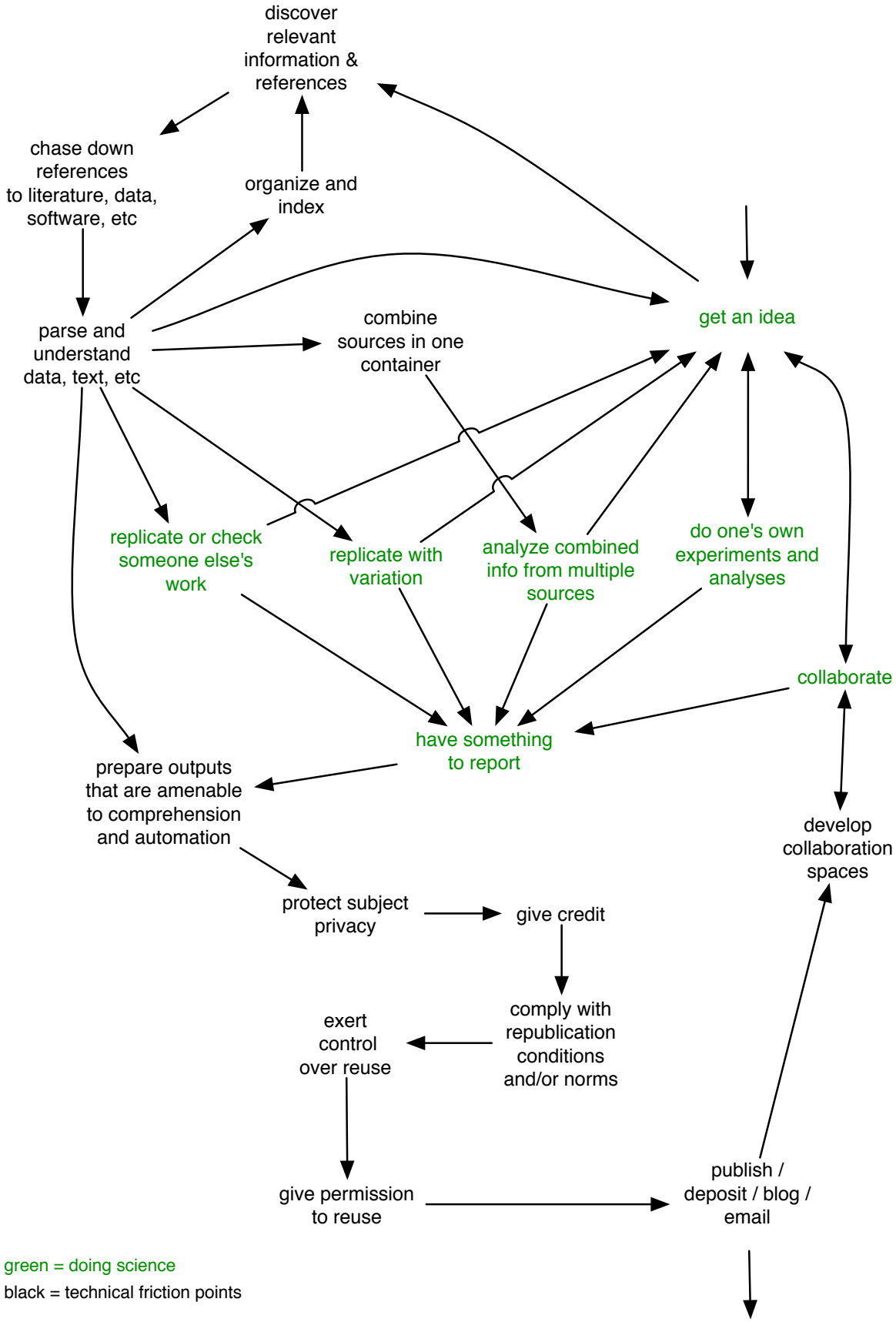
Because scientific research is costly, the efficiency with which it is carried out is a matter of great importance. Not only do interventions that lower transactions overall have the potential for a high return on investment, but efficiency itself creates new research opportunities. Any particular improvement to a resource (report, data set, etc.) is seen by all downstream transactions that involve it, and techniques for improving efficiency can themselves be reused across many resources. Investments that coordinate previously unstandardized practices may lead to network effects, in which the success of a coordinated effort leads to voluntary further adoption of a standard.

When a resource such as a data set is difficult to interpret or (re)use, someone may choose to forego reuse entirely and instead re-create it by repeating an experiment or analysis. This is a particularly egregious example of high transaction cost. It would be wise to look at ways in which such expensive redundancy can be avoided.

A complete survey of opportunities for improving transaction efficiency, even if it were possible, would be tedious. Instead this report touches on a range of phenomena brought to the author’s attention in the course of observing and prototyping open science over the past six years. Take this document like a shower, not like a bath. It is meant as a starting point for discussion of potential intervention points. The best opportunities may well fall outside the coverage of this report.

Friction points

The following diagram is a cartoon showing activities that investigators engage in during the course of a study. Some activities relate directly to scientific study, while others involve communication with other investigators - consuming others' information, or generating new ones. Technology dominates information-pushing, so each communication related activity is a potential friction point in the research process. The less time is spent on information wrangling, the more time can be spent working directly on scientific problems.


discover relevant information & references
chase down references to literature, data, software, etc
organize and index
parse and understand data, text, etc
combine sources in one container
get an idea
replicate or check someone else's work
replicate with variation
analyze combined info from multiple sources
do one's own experiments and analyses
collaborate
have something to report
prepare outputs that are amenable to comprehension and automation
develop collaboration spaces
protect subject privacy
give credit
comply with republication conditions and/or norms
exert control over reuse
give permission to reuse
publish / deposit / blog / email
green = doing science
black = technical friction points

Research mechanics

The upper part of the diagram shows tasks involved in consuming resources developed by others.

1. Discovering relevant resources: There is a huge variety of resources available for discovery, ranging from general-purpose search (e.g. Google) to nationally supported discipline-oriented databases (e.g. the Mouse Genome Informatics database), and “boutique” databases, (e.g. UbiProt, collating protein substrates of ubiquitylation) Pain points include search precision - finding what you need, and only that - and meta-discovery or being aware of search services that can help.
2. Chasing references: The links that tie one resource to another, e.g. a research report formally citing another report or a results table linking to the software used to generate it, are the bread and butter of scholarly literature. The transition from old fashioned references to machine-actionable ones (URIs, including DOIs) is often incomplete. Links fail for a number of reasons, with link stability and content persistence often being quite difficult to achieve.
3. Parsing a resource: To use a data set, image, or report, one must have software that parses it, i.e. takes it apart into its component pieces for processing. We generally take this for granted since we have Excel for parsing Excel files, web browsers for HTML, image viewers for images, compilers for software, and so on. But the durability of the software needed to process particular resources is not assured, and stock parsers for file formats do not always suit automation needs - for example they are often tied to a user interface and must therefore be used under human supervision. Obscure legacy data formats are a continuing challenge.
4. Understanding in depth: Parsing is the easiest part of reuse - figuring out what something means, in scientific terms, can be an enormous challenge. Numbers are not numbers; they might be measurements or calculations, and to understand them requires knowledge of units, protocols, and provenance (how they come to be where they are). Species names and gene symbols can only really be understood with knowledge of which authority was used to connect the name with some description. "Sample weight" might include the container in which it was collected or not.
5. Syntactic integration: One combines information sources either to enlarge a pool of similar information records, or to correlate one kind of information against another that shares particular data elements but not others. When information is combined from multiple sources the stakes go up for both syntactic and semantic processing. On the syntactic side, a common platform has to be established in which sources can be compared or combined. The two sources may start in incompatible formats, and one of the two, or a third, has to be chosen as the "hub" format to be used for integration. We see Excel, relational databases, and more recently RDF used for this purpose.
6. Semantic integration: The simplest kind of conversion is unit conversion; "miles" in one data set might be nautical while "miles" in another could be statute miles, and a conversion factor has to be applied to align the two sources. All the problems experienced in search - is this the same as that? how do these compare? - are found here, often with much less contextual information to go by.
7. Processing text: There is no end to the variety of text processing tasks one would want to do - applied either to a few dozen articles listed in one's Mendeley reference list, to millions found in PubMed. One looks for gene or protein names, drug names, alleles, processes, reagents, or any other entity of interest; with more ambition one might try to

find relationships between entities such as the co-occurrence of a NSAID therapeutic and an alkali metal in the same paragraph. You might attempt such searches on the fly, or perform a large number of searches all at once and save the results for quick access.

8. Building an index or knowledge base: Most knowledge bases probably begin life as single investigator projects, where a need has arisen for keeping track of specific information in order to advance that investigator's research interests. A typical example might be a phosphorylation site knowledge base. Initially created for a particular purpose, such efforts often bloom into institutional and/or collaborative projects. These can eventually be important resources for large numbers of scientists.
9. Collaboration: It is useful for investigators to work together in a variety of ways. Communicating and working closely in person, by email, etc. is more efficient than individual efforts coupled only through the peer review system. Technology has come to figure prominently in collaboration, providing not just communications technology but virtual online workspaces to keep track of joint activities and results. Any improvement in these technologies is going to yield an improvement in science.

Dissemination mechanics

1. Protecting subjects' privacy: This is a very active area of research and debate. We know that de-identification rarely if ever is truly effective, making the creation and maintenance of "data sharing clubs" allowing limited circulation of partially de-identified data a priority. Informed consent to data disclosure might provide one way to open up data that would otherwise have to be locked up in a sharing club.
2. Giving credit: Reference and credit are conflated in practice, so the way to express a debt is to cite something someone has produced. This puts pressure on the mechanism of reference and suggests that all relevant resources - notably data, ontologies, and software - should have formal references, not just casual mentions in text. One way to establish first-class references to resources is to dress them up as research reports.
3. Complying with license conditions and norms: Data gets processed and reprocessed through generations of research studies. If data sets are protected by copyright or sui generis database rights (or contract) and then licensed with conditions, those conditions must be respected as derived information moves downstream. This problem has been called "attribution stacking" but other conditions, such as the obligation to report details of the manner of adaptation, also "stack". There may be technical solutions to make this manageable but as this is not yet experienced by the community as a pain point, the economies of scale require to develop a solution have not been realized. It would be useful to explore the nature of the difficulty before it does.
4. Giving and controlling permission to reuse: Scientists might feel they need need certain kinds of exclusion before they share, and may decide not to share if these needs will not be met through legal means. Two concerns often heard are protection against threats to integrity - they don't want their data misrepresented, or changed and corrupted in some way - and competition in commercial use of the resource. Do these problems have solutions with technical components? In addition a scientist might want to share data, software, etc. but not know how technically to signal this - where do you put license metadata in an RDF file, or an Excel file, such that it will be found both by people and by automation?

Automation reduces friction

The common thread here obviously is automation. The way to reduce technical (as opposed to legal or social) friction is through automation, and automation can only work reliably when inputs and outputs are amenable to automation. In most cases this implies that meaning must be known and expressed in a language that can be understood (to some extent) by a computer. Recording meaning and re-expressing it may require manual work - for example, selection of the correct ontology term for a unit of measurement or for a drug, or canonical identifier for a gene or periodical. But once the work is done, the benefits can be reaped: canonical, meaningful expression that avoids synonymy and polysemy pitfalls.

It is these considerations that led the Science Commons data project to focus on developing and promoting technologies for identifiers, ontologies, and study and resource descriptions. To automate activities that are relevant to scientific investigation requires this.

Identifiers and ontology terms are used as part of a language, and this is where RDF (a resource description framework), OWL (an ontology language), and machine inference come in. These have been associated with the technical-sounding phrase "knowledge representation" which has its roots in artificial intelligence, which would seem remote from the mundane considerations of this report. However when it is understood that the game is not one of "representing" (whatever that is) but *expressing* information in a form amenable to machine manipulation, or *translating* information from human-ese to computer-ese, the function of the activity becomes apparent, and the need seems inevitable.

RDF is also associated with the phrases "semantic web" and "linked data" which are red herrings in this discussion - RDF makes perfect sense without these ideas. The important idea to take from "semantic web" is simply that RDF, like HTML, lets the composer of a resource use hyperlinks to help lead a reader to documentation for elements of expression. One would hope this is true of any format used for expression scientific information.

Because computers are so central to scientific work, and hold such promise for improving efficiency, what we should call for is a complete reorientation of scientific communication toward computational reproducibility and clarity. The following examples illustrate some aspects of what this reorientation might entail.

Example - computational reproducibility

As an example illustrating the gap between sensible (low-friction) standards of computational reproducibility and practice, consider the story behind a computational methods manuscript, for which one of us at Science Commons served as editor. When the manuscript was originally submitted, it was impossible to reproduce the computational result, in spite of software being provided, because not enough detail was provided on how exactly to run the software to obtain the reported results. Instructions provided on request did not initially work due to lack of detail. The final outcome only yielded computational reproducibility because of the particularly insistent and involved editor.

The editor had to report bugs in the software and in the execution instructions to the authors, wait for fixes, and iterate until the instructions worked for him. But this is unusual. There is no place in the current editorial process for ensuring that things like this work, and no author incentives for this kind of work.

> Supporting Information S2. Source document extracts and scripts. The results described in this paper can be recreated by following the workflow described in the file entitled "Instructions for Executing T4SS Named Entity Recognition Workflow.docx."

This example is informative because it is typical of what is needed to improve overall system efficiency in many different scenarios. The editor, acting as a reader advocate, seemed to create additional friction for the authors that another editor might not have. But suppose that he had not. With work and a good deal of reverse engineering, a reader might have been able to reproduce the reported results. More likely they would have contacted the author for instructions and iterated just as the editor did. The amount of work for the authors would have been the same in the end. But by putting reproducibility instructions in the supplementary materials, any number of readers are able to reproduce the results without interacting with the authors - clearly a more reliable, scalable, and efficient result.

Example - syntactic integration

Another interesting example is the "Debugging the bug" exercise which brought two independently developed metabolic pathway networks for *E. coli* into comparable form so that each could be checked against the other. In principle the two networks should be in agreement wherever they overlap. Because both studies involve the same strain, any incompatible differences reflect quality issues with either one network or the other.

The purpose of the exercise is to reconcile the two networks, and the first step is to put them in a common syntactic format so that semantics can be compared without distraction. The inputs start out in incompatible file formats, and one approach would have been to convert one of them into the format used by the other. Instead, however, a more general third "hub" format was chosen, and the two inputs converted into it, to take advantage of analysis tools operating on the hub format.

The semantic analysis uncovered large number of mistakes in both networks, and provided feedback that was used for strengthening both.

There are two lessons here. Syntactic conversion is usually fairly easy, and often is only the beginning of any combined analysis. It is easy to confuse syntactic integration with semantic integration; any claim that a tool capable of the former is also doing the latter must be met with skepticism. On the other hand, if the two inputs had been provided in a standard format in the first place, then the conversion steps (often a source of both inefficiency and error) would have been unnecessary, and the semantic gaps might have been more evident at an earlier point.

Example - semantic integration

The US National Center for Biotechnology Information (NCBI) runs a resource call dbGaP that pools interrelated genotype and phenotype (disease and clinical variables) data sets. If we look at the example "NHLBI Coronary Artery Risk Development in Young Adults (CARDIA) Candidate Gene Association Resource (CARe)" we find that total cholesterol was measured in that study:

> Variable Name and Accession
> Variable Name: AL1CHOL
> Variable Accession: phv00113700.v2.p2

Variable belongs to dataset: pht001588.v2.p2 : A4LIP: 1985-1986, Year 0. Lipoproteins and Triglycerides. Cardiovascular disease and its risk factors among young adult participants.
Variable Description
TOTAL CHOLESTEROL (MG/DL)

The variable name AL1CHOL is a label specific to this data set. Obviously other data sets have measured total cholesterol in the same way, and relating all such studies is scientifically interesting (show me cardiovascular disease studies in which total cholesterol was measured, etc.). The dbGaP curators could have integrated the data sets along this axis by using text strings or text mining, but this would have led to many false positives (variables described the same way but differing in some important regard) and false negatives (the same variable described in ways not recognized as the same). They are to be lauded for their integrity, but it's hard to avoid thinking an opportunity's being missed here.

There is huge opportunity here, with thousands of variables across thousands of data sets to be unified. Unfortunately, if this is done retrospectively by the dbGaP curators, each case would need to be vetted by curators. This is work that is done more efficiently by the data set's authors or publisher, since they are familiar with the study and can make use of information that might not have been captured in the data set or its associated documentation. It is work (friction), however, and we do not have systems in place for holding authors accountable for producing machine-readable documentation. The work therefore has to be done downstream instead.

Discussion

The effort to build an effective research commons needs to incorporate the scientific legacy and plan for the future. Therefore some opportunities are prospective (they support resource producers to make things better by promoting changes to practice that make use easier for consumers), while others are retrospective (further invest in tools to integrate and make usable our legacy artifacts).

We can imagine a number of activities, support for which will help to address the technical friction points we have identified. While much work towards this end is proceeding in the community, it is our assessment that this work requires significant additional coordinated effort. Therefore suggestions range from tactical to strategic.

- Support basic research to understand and improve common uses of research resources: Study fundamental logical and social principles of scientific practice, such as information economics and incentive structures
- Support and nurture service organizations that share this vision: Coordination groups, help desks, efforts to improve repository interoperability (e.g. OBO Foundry)
- Support and engage efforts to standardize: participate in standardization processes and outreach, perhaps through established organizations such as W3C, NISO, FGED, TDWG
- Standardize and coordinate: participate in standardization processes and outreach, perhaps through established organizations such as W3C, NISO, FGED, TDWG
- Support the technology that *has* been standardized: E.g. better user experiences for semantic web based technologies, including inference

- Create examplars: Focus on one investigator, project, or discipline, pursue technical interoperability pilot projects and widely disseminate those results.
- Educate: Provide tutorials and courses aimed at teaching scientists good data and software stewardship, at both the technical and organizational level.

What makes the task of fixing friction points in scientific communication so challenging is that by their nature any change requires coordination between "senders" (publishers, however informal) and "receivers" or readers. There is no point in asking senders to change if there is no demand on the part of receivers. Conversely, receivers won't be motivated to build or learn tools that process better outputs if there are no senders producing them.

Replicating solutions that have improved our ability to share and remix cultural work with less friction form the basis of work that will be necessary to create a research commons. But building an effective, productive, and efficient commons to support science and deliver consequent benefits to society will need new approaches. Taking the colossal factory of scientific research and retooling for computational tractability requires a new branch of engineering. To bring this about one needs cooperation among a wide range of specialties:

- Practitioners of science: Scientists, researchers, graduate students, laboratory support
- Science-oriented technology pioneers - to show what's possible and what's needed - Rod Page is a good example
- Infrastructure coordinators - to know the communities' needs and make computational connections between projects and recognize where more work needs to be done
- Standardization enthusiasts - to prove commitment, legitimize good ideas, and document ideas so that they can be disseminated to wide audiences,
- Community organizers - to raise awareness and recruit followers.
- Educational and research institutions - to teach and encourage a new generation as they grow up doing research

What kinds of efforts are likely to scale, i.e. to reap benefits across many disciplines, institutions, nations without constant heavy investment? Precedents include the development and propagation of computer science into mainstream academia, development and adoption of the Web, the spread of GNU/Linux, participation in Facebook, the broad availability, quality and scope of PubMed.

As Steve Jobs often pointed out, ideas that scale are not unexpected; not the outcome of market research; not a committee effort; a clearly good idea only after the fact (i.e. risky); they induce network effects (utility per user increases with number of users); and in most cases their infrastructure is replicable (legally, technically, administratively).

Collectives potentially wield influence and purchase power that individuals cannot, but scientists, as a community, are notoriously poorly organized. On the other hand top down efforts have a mixed track record. The trick here is finding the proper middle ground between anarchic individual (in)action on the one hand, and bureaucratic outsider efforts on the other.

Open access and open data form the basis of open science, but are in themselves not adequate enablers for open science. The problems of open science are much deeper and pervasive than the legal mechanics, which don't even get the notice of most scientists.

Acknowledgments

Alan Ruttenberg and the author worked together at Science Commons building systems and learning first hand what's hard about building an effective research commons. This document is the product of our shared experience. John Wilbanks had the vision to support and encourage the work that led to this report. Many thanks to Alan Ruttenberg, MacKenzie Smith and David Kindler for help in preparing this report.

Further reading

1. Philip Bourne. What Do I Want from the Publisher of the Future?http://dx.doi.org/10.1371/journal.pcbi.1000787
2. Common Crawl. http://www.commoncrawl.org/
3. dbGaP. http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap
4. Datacite. http://datacite.org/
5. Debugging the bug. http://bio.freelogy.org/wiki/Debugging_the_bug
6. Data Privacy Lab: Proposed Changes to the Common Rule. http://dataprivacylab.org/projects/irb/index.html
7. Data reuse Mendeley group. http://www.mendeley.com/groups/544251/data-reuse/papers/
8. Filtered Push. http://etaxonomy.org/mw/FilteredPush
9. Francesca di Massimo. How to make the most of the open data opportunity. Microsoft EMEA Press Centre, June 2011. http://www.microsoft.com/presspass/emea/presscentre/pressreleases/June2011/MakeTheMostOfTheOpenDataOpportunity.mspx "The prevalence of open data is published to application-specific, non-reusable formats that also lack terminology and data consistency."
10. G2D: Candidate genes to inherited diseases. http://www.ogic.ca/projects/g2d_2/
11. Tom Howard et al. Opportunities for and Barriers to Engineering Research Data Re-use. ERIM, University of Bath, 2010. http://opus.bath.ac.uk/21166/1/erim3res100805tjh10.pdf
12. Adrian Johns. *The Nature of the Book*. University of Chicago Press, 1998.
13. myExperiment. http://www.myexperiment.org/
14. The Ontology for Biomedical Investigations (OBI). http://obi-ontology.org/page/Main_Page
15. The Open Biological and Biomedical Ontologies (OBO). http://www.obofoundry.org/
16. Smith B, et al. "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration", Nature Biotechnology 25, 1251 - 1255.
17. Open PHACTS: Open Pharmacological Space. http://www.openphacts.org/
18. ORCID: Open Researcher and Contributor ID. http://orcid.org/
19. Pubby: A Linked Data Frontend for SPARQL Endpoints. http://www4.wiwiss.fu-berlin.de/pubby/
20. Ray Bat. http://raybat.org/
21. Jonathan Rees. Recommendations for independent scholarly publication of data sets. Creative Commons Working Paper, 2010.
22. Relfinder. http://www.visualdataweb.org/relfinder.php
23. Science Collaboration Framework (SCF). http://sciencecollaboration.org/
24. Shared Names project. http://sharedname.org/
25. Semantic Web Applications in Neuroscience. http://swan.mindinformatics.org/
26. Suzanne Little. Survey of Metadata Registries. DART Project, 2006.
27. Taverna. http://www.taverna.org.uk/
28. Wikipedia article on interoperability. http://en.wikipedia.org/wiki/Interoperability